Los Alamos
NATIONAL LABORATORY
EST 1943

Los Alamos releases file index product to software community

March 22, 2018

Grand Unified File Index (GUFI) hits GitHub for users

LOS ALAMOS, N.M., March 22, 2018—Resolving the supercomputer challenge of searching and retrieving files could now be far simpler, with a tool developed by Los Alamos National Laboratory and released today to the GitHub open-source software site. The Grand Unified File Index (GUFI) is designed using a new, heirarchical approach to storing file metada, allowing rapid parallel searches across many internal databases. Queries that would previously have taken hours or days can now be run in seconds.

“We anticipate that the Grand Unified File Index will have a big impact on the ability for many levels of users to search data and get a fast response,” said Gary Grider, division leader for High Performance Computing at Los Alamos. “Compared with other methods, the Grand Unified File Index has the advantages of not requiring the system administrator to do the query, and it honors the user access controls allowing users and admins to use the same indexing system,” he said.

Why develop a new search-and-retrieval tool? At Los Alamos and other supercomputing facilities around the world, databases for file metadata may potentially hold hundreds of millions of records, yet they are typically inefficient for the kinds of searches that are actually needed.

In recent decades, a major issue has been providing storage that could handle huge flows of data going in and out of state-of-the-art supercomputers. Handling these huge volumes quickly and economically allows the machines to make progress on scientific calculations that support national security, as well as for basic scientific research in fields such as engineered materials, biological processes, and earth systems modeling.

One important solution to the storage bottleneck has been the Parallel File System (PFS). A PFS allows many related streams of data to be moved at the same time, without losing track of how they are related. Unfortunately, searching through the lists of files that are stored in such systems remains difficult.

“Simple queries, such as ‘where is the simulation data that was done with that new computational method?’ could bring the PFS to its knees,” said Jeff Inman, one of the developers of the new search tool. System administrators frequently need to query to find which files should be archived, who is using the most storage, whether a dataset has been moved from PFS to tape, and so forth, he noted, and any of these queries may seriously compromise the performance of the file-system to send and receive data.

The trick used by GUFI is to store file-metadata in a hierarchy of databases, matching the hierarchy of folders. This allows rapid parallel searches across many databases, and allows access-permissions to be managed in the same way they are managed in a normal hierarchy of folders. GUFI can hold file-metadata from tape-archives, PFS, and other kinds of file-systems, unifying information from all the places where a file might reside.

The Laboratory is planning to initially present the work at a Microsoft gathering in March and a subsequent HPE session, then rolling it out at the IEEE Massive Storage Systems and Technologies Conference (MSST) May 14.

GUFI is now available at https://github.com/mar-file-system/GUFI.git as open-source software for interested users to download and explore.

Funding: The GUFI project was funded by the U.S. Department of Energy’s National Nuclear Security Administration Advanced Simulation and Computing Program.